TOWARD *De Novo* PROTEIN DESIGN FROM NATURAL LANGUAGE

Fengyuan Dai[1], Yuliang Fan[1], Jin Su[1], Chentong Wang[1], Chenchen Han[1], Xibin Zhou[1], Jianming Liu[1], Hui Qian[1], Shunzhi Wang[2], Anping Zeng[1], Yajie Wang[1] and Fajie Yuan[1]*

[1]Westlake University, [2]University of Washington

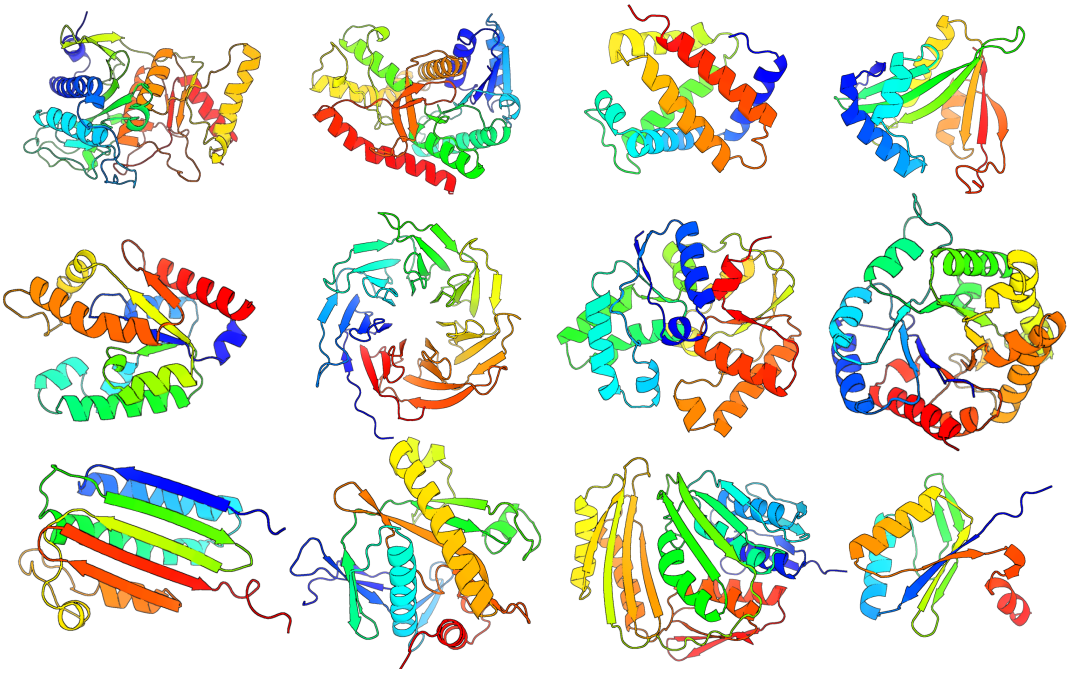

Figure 1: **Showcase of designed proteins.** Refer to Appendix A for a complete list of prompts. Pinal can generate diverse proteins from both short or long protein language descriptions.

ABSTRACT

De novo protein design (DNPD) aims to create new protein sequences from scratch, without relying on existing protein templates. However, current deep learning-based DNPD approaches are often limited by their focus on specific or narrowly defined protein designs, restricting broader exploration and the discovery of diverse, functional proteins. To address this issue, we introduce Pinal, a probabilistic sampling method that generates **p**rote**i**n sequences using rich **na**tural **l**anguage as guidance. Unlike end-to-end text-to-sequence generation approaches, we employ a two-stage generative process. Initially, we generate structures based on given language instructions, followed by designing sequences conditioned on both the structure and the language. This approach facilitates searching within the smaller structure space rather than the vast sequence space. Experiments demonstrate that Pinal outperforms existing models, including the concurrent work ESM3, and can generalize to novel protein structures outside the training distribution when provided with appropriate instructions. This work aims to aid the biological community by advancing the design of novel proteins, and our code will be made publicly available soon.

*Corresponding author: yuanfajie@westlake.edu.cn.

1 INTRODUCTION

Proteins are fundamental to life, playing critical roles in biological processes across all living organisms. Protein design aims to customize proteins for specific biological or biomedical purposes. Traditional protein design methods (Dougherty & Arnold, 2009; Accuracy, 2003), while effective, are often limited by their reliance on existing protein templates and natural evolutionary constraints. In contrast, *de novo* design (Huang et al., 2016) benefits from the two perspectives. Firstly, nature has only explored a small subset of the possible protein landscape. Secondly, the biological attributes selected by evolution may not align with our specific functional requirements. *De novo* design allows us to create entirely new proteins with desirable structures and functions, thus overcoming the limitations of traditional methods.

Although *de novo* protein design (DNPD) using deep learning (Watson et al., 2023; Ingraham et al., 2023; Krishna et al., 2024) has gained considerable attention, current methods often operate under rather limited conditions. These methods typically focus on either unconditional design (Lin & AlQuraishi, 2023) or on specific functions such as conditioning on control tags, motif scaffolding, and binder design (Madani et al., 2023; Watson et al., 2023). Given the versatile functions and biological significance of proteins, these approaches provide only a limited view of the target protein and may not fully capture their complexity and diversity. To this end, we propose a more ambitious and general approach: designing *de novo* proteins from natural language. This method leverages the descriptive power and flexibility of natural language to accurately communicate design objectives and functionality requirements to the protein generator.

Protein molecules exhibit a profound relationship between their structure and function (Dill & MacCallum, 2012). Inspired by traditional physics-based approaches (Cao et al., 2022), we propose an intuitive method: rather than designing protein sequences directly from natural language descriptions of function, we first translate these descriptions into structural information and then generate sequences conditioned on both the language description and its structure. To achieve this, we first employ an encoder-decoder architecture named T2struct, which is designed to interpret natural language and derive structural information from it. Instead of dealing with explicit 3D structures, we use discrete structure tokens generated by the vector quantization technique (van Kempen et al., 2022), which have been shown to offer better scalability for larger datasets (Su et al., 2024a; Hayes et al., 2024). Subsequently, we modify and retrain SaProt (Su et al.), a structure-aware protein language model, referred to as SaProt-T, to understand natural language inputs and enable sequence design based on the given backbone and language instructions. This pipeline provides an effective pathway to map natural language to protein sequences. It ensures that the designed proteins accurately express the desired functions and exhibit robust foldability.

To summarize, this work makes three key contributions:

- We argue and provide evidence for the first time that end-to-end training for text-to-sequence mapping is difficult due to the vast expanse of the protein sequence space. Instead, a two-stage design approach leverages the smaller structure space to impose greater constraints on sequence generation, resulting in significantly higher performance across various protein design metrics.

- We implement this insight using deep learning and propose a novel DNPD pipeline called Pinal. Pinal consists of two network components, T2struct and SaProt-T, which handle the translation of natural language into protein structure and the subsequent sequence design, respectively. Additionally, we derived an optimal sampling strategy to seamlessly integrate T2struct and SaProt-T into the Pinal framework.

- We systematically validate Pinal through a series of experiments under both concise and detailed instructional scenarios. Pinal consistently outperforms other models, including prior methods and the concurrent ESM3 (Hayes et al., 2024). Furthermore, we find that Pinal demonstrates strong generalization capabilities, enabling it to design novel proteins beyond the training distribution.

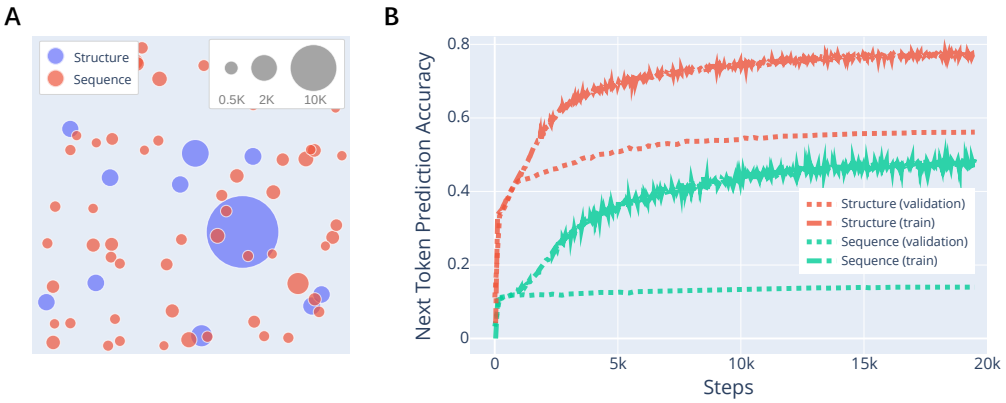

Figure 2: **Learning to predict structure is easier. A**: Visualization of protein sequence and structure space using Swiss-Prot database as an example. Each circle represents a cluster, and the size of the circle corresponds to the number of proteins within that cluster. 30% token identity is used for clustering, resulting in 39,322 structure clusters and 126,185 sequence clusters. We only plot the largest clusters for brevity, while ensuring that the total number of proteins in the displayed structure and sequence clusters remains the same. **B**: Next token prediction accuracy trend training on structure and sequence token. The training and validation sets are split by structural similarity to ensure that similar proteins do not overlap between the two sets.

2 Avoiding Designing Sequence Directly

The most intuitive approach for DNPD from natural language is to use an encoder-decoder architecture. In this setup, the encoder represents the natural language input, while the decoder generates the corresponding protein sequence. We refer to this method as end-to-end training. However, this approach can be challenging due to the vast expanse of the protein sequence space, which makes accurate sequence generation difficult. In contrast, protein structure is more conserved and intuitively much easier to predict and generate.

In Fig.2A, we visualize both the protein sequence and structure space of the Swiss-Prot database. The protein structure here is represented by discrete structural tokens via Foldseek(van Kempen et al., 2022), which has the same number of alphabet as that of amino acids, i.e., 20. As clearly shown, the protein sequence space is much larger and more diverse than the structure space.

In Fig. 2 B, We apply the same encoder-decoder architecture to train both a language-to-sequence model and a language-to-structure model, with protein structures still represented by Foldseek tokens. As expected, deep learning models that learn the language-to-sequence space find it more challenging to achieve ideal next-token prediction accuracy compared to the language-to-structure model in the validation set. This finding also explains why previous language-guided protein design models, such as ProteinDT (Liu et al., 2023), exhibit poorer performance, as discussed in Section 4.2. In that section, we also conducted a direct comparison of end-to-end training (i.e., ProGen2-ft) with our proposed two-stage model, Pinal, which further supports this argument.

3 Pinal for Language-Guided DNPD

3.1 Pinal Framework

Motivated by the above analysis, we decompose the protein design process into two primary steps: first, translating natural language descriptions of protein into structures, and then generating protein sequences based on both the structure and language description. Specifically, as the structure is determined by sequence, we modify the probability of protein sequence s to the joint distribution of s and additional protein structure information c:

$$p(s \mid t) = p(s, c \mid t),$$

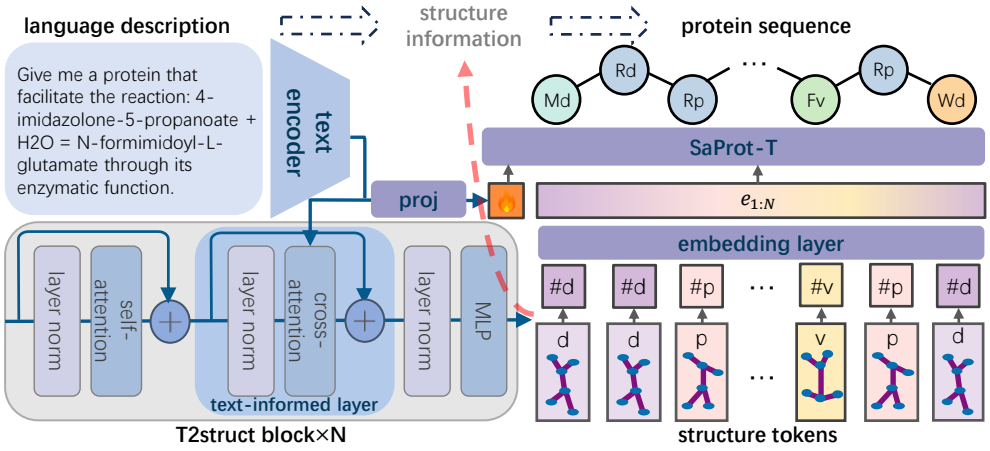

Figure 3: **Overview of Pinal.** The text encoder first encodes the provided protein description into textual embeddings. T2struct then uses an encoder-decoder architecture to predict discrete structural tokens from these embeddings. These structural tokens are subsequently concatenated with the projected textual embeddings and fed into SaProt-T to facilitate the design of a *de novo* protein sequence.

where t denotes the language descriptions. According to the Bayes' theorem, the above joint distribution can be formulated as:

$$p(s, c \mid t) = p(c \mid t)p(s \mid c, t). \tag{1}$$

On the right side of the equation, $p(c \mid t)$ indicates the probability of the structural information aligning with t, and $p(s \mid c, t)$ implies the probability of a sequence capable of folding into the depicted structure and expressing the desired function as specified by the given language description. In practice, we leverage T2struct to predict $p(c \mid t)$, and SaProt-T to predict $p(s \mid c, t)$, thereby generating protein sequences that account for both structural folding and functional expression, as illustrated in Figure 3.

3.2 Decoding Structural Information from Natural Language

Protein structures can be represented in various ways, including explicit 3D structures and discrete structural tokens (van Kempen et al., 2022; Lin et al., 2023; Gao et al., 2024). However, our Pinal framework is better suited to the latter approach, given its sampler design in Section 3.4. In this paper, we use 3Di token sequence to represent structures generated by Foldseek (van Kempen et al., 2022).

To represent $p(c \mid t)$, we model the conditional probability of structural tokens c in an auto-regressive manner:

$$p(c \mid t) = p_{\theta_1}(c_{1:N} \mid t) = \prod_{n=1}^{N} p_{\theta_1}(c_n \mid c_{<n}, t), \tag{2}$$

where θ_1 denotes the learned parameter. To generate structural tokens conditioned on text embeddings, we employ an encoder-decoder architecture (Vaswani et al., 2017). Specifically, we leverage the pre-trained text encoder from PubMedBERT (Gu et al., 2021) (109M) as the language encoder. For the 3Di structural token decoder, we utilize a randomly initialized GPT-2 architecture (Radford et al., 2019) with 3Di token embeddings, enhanced by a text-informed layer in each block (114M). This text-informed layer incorporates layer normalization, a cross-attention mechanism, and a residual connection.

3.3 Sequence Generation from Structural and Natural Language Condition

Given a sequence of structural tokens $c_{1:N}$, SaProt (Su et al., 2024a) predicts the corresponding amino acid sequence. Specifically, each structural token c_i is paired with a masked amino acid $\#$, represented as $\#c_i$. SaProt takes structural sequence $(\#c_1, \#c_2, ..., \#c_N)$ as input and outputs the

structure-aware sequence $(x_1, x_2, ..., x_N)$, where x_n denotes the combination of the amino acid and the structural token, *i.e.* $s_n c_n$. Although SaProt demonstrates impressive performance in predicting sequences based on structural tokens, its predicted sequences are not explicitly conditioned on textual descriptions.

To model $p(s \mid c, t)$, we re-train SaProt with text as additional input, referred to as SaProt-T. Given the textual embeddings after the pooling layer, $e_t \in \mathbb{R}^{1\times d_t}$, we project them using a trainable matrix $W \in \mathbb{R}^{d_t \times d_s}$, where d_t and d_s represent the embedding dimensions of the text encoder and SaProt-T, respectively. The resulting embeddings e_{input} are concatenated with embeddings of the structural token sequence $e_{1:N} \in \mathbb{R}^{N\times d_s}$:

$$e_{input} = [e_t \times W, e_{1:N}].$$

SaProt-T takes e_{input} as input and is trained to predict the masked amino acid at each position. Denoting θ_2 as the parameter of SaProt-T, we calculate the product of conditional probabilities over the length of the protein as follows:

$$p(s \mid c, t) = \prod_{n=1}^{N} p_{\theta_2}(s_n \mid c, t) = \prod_{n=1}^{N} p_{\theta_2}(x_n \mid e_{input}). \tag{3}$$

3.4 Sampling in the View of Probability

While designing the protein backbone structure based on its function, followed by designing the sequence based on this structure, is a commonly used heuristic in traditional protein design, this paper introduces a more rigorous mathematical explanation for this from a Bayesian perspective (i.e., Eq.1). More importantly, we derive an optimal sampling scheme for the entire two-stage generation process, as outlined below.

$$\begin{aligned} \arg\max_s p(s \mid t) &= \arg\max_s \log p(s \mid t) = \arg\max_s (\log p(c \mid t) + \log p(s \mid c, t)) \\ &= \arg\max_s \sum_{n=1}^{N} \Big[\log p_{\theta_1}(c_n \mid c_{<n}, t) + \log p_{\theta_2}(s_n \mid c, t) \Big]. \end{aligned} \tag{4}$$

Note: $$\arg\max_s (\log p(c \mid t) + \log p(s \mid c, t)) \neq \arg\max_c \log p(c \mid t) + \arg\max_s \log p(s \mid c, t) \tag{5}$$

The above equation illustrates how the probability of a protein sequence can be precisely estimated using the Pinal pipeline by leveraging T2struct and SaProt-T. Instead of sequentially determining the most aligned structure and then the optimal sequence , as shown on the right side of Eq. 5—an approach that often leads to suboptimal outcomes (see Fig.6B)—we consider the joint probability distribution of structure and sequence to achieve the optimal protein sequence generation by computing Eq. 4. In practice, given the challenge of exploring the entire range of s values, we limit our exploration to K sampled values. That is, we first sample K structural sequences using multinomial sampling via T2struct and then apply a greedy search to determine the corresponding protein sequence via SaProt-T. After that, we compute Eq. 4 for these K candidates and select the top candidates. We set K to 50 throughout this paper (see Section 4.5 for ablation).

4 Experiments

4.1 Experiment Setup

Dataset Construction. We use the subset of the dataset from (Su et al., 2024b). The dataset is sourced from the human-reviewed Swiss-Prot database (Boeckmann et al., 2003), where descriptive protein information is categorized into sequence-level and residue-level details. It contains 560K protein sequences and 14M protein-function pairs.

Notably, unlike the common practice of splitting data based on sequence similarity, we tackle a more challenging scenario by splitting data based on structural similarity. This approach is motivated by the fact that protein structures are typically more conserved than protein sequences, meaning that dissimilar sequences might still adopt similar structures. This ensures that the model's generalization can be properly assessed. Following (Kucera et al., 2024), we conduct structural clustering using Foldseek (van Kempen et al., 2022) and define structural similarity based on the Local Distance

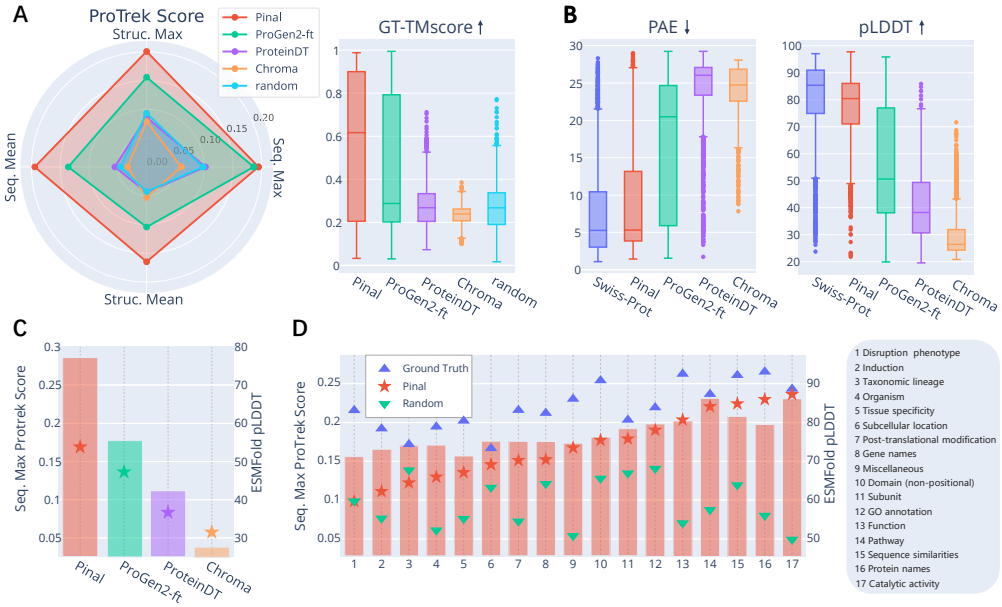

Figure 4: **Assessment of designed protein from language descriptions. A**: Alignment of designed proteins with given language descriptions, assessed by `ProTrek Score` and `GT-TMscore`. **B**: ESMFold evaluation of designed proteins foldability. **C**: Evaluation of protein design efficacy based on short prompts. **D**: Detailed performance analysis across diverse protein function descriptions. In C and D, the bar represents `pLDDT` and the star represents `ProTrek score`.

Difference Test (LDDT) metric. Specifically, the LDDT threshold is set to 70%, resulting in 57K clusters. For validation, we allocate 300 clusters (3481 proteins), while an additional 50 clusters (307 proteins) are set aside for testing. The remaining data constitute the training set.

Baselines. We compare Pinal with two baseline methods that support protein design with textual description, namely ProteinDT (Liu et al., 2023) and Chroma (Ingraham et al., 2023). ProteinDT is a multi-modal framework that integrates text with protein design, employing the architecture of DALL·E 2 (Ramesh et al., 2022). We use its auto-regressive version[1], which has been reported to exhibit better performance. Chroma is a diffusion model that can directly sample protein structures and sequences guided by a pre-trained protein captioning model.

In addition to these methods, we train another end-to-end models for a more nuanced comparison. We use pre-trained ProGen2 (Nijkamp et al., 2023) as the decoder and PubMedBERT as the encoder, denoted as ProGen2-ft (271M). This model are trained until convergence with a batch size of 768 on the aforementioned dataset. Lastly, we compare Pinal with a concurrent work, ESM3, a frontier model trained with functions, protein sequences, and structures simultaneously. Note that ESM3 can only support keywords rather than detailed text descriptions.

Metrics. We evaluate proteins from two perspectives: foldability and language alignment. **For foldability**, we analyze the designed sequences using the single-sequence structure prediction model, *i.e.* ESMFold (Lin et al., 2022). We compute the average predicted local distance difference test (`pLDDT`) (Wang et al., 2024c) and predicted aligned error (`PAE`) across the whole structure according to the output of ESMFold. It's important to note that `pLDDT` above 70 and `PAE` below 10 are commonly used thresholds indicating high prediction confidence. These thresholds suggest a high probability that the predicted protein sequences can fold into the corresponding predicted structures. **For language alignment**, we measure the structural similarity between the ESMFold-predicted structure of the designed sequence and the ground truth using TMscore (Zhang & Skolnick, 2004) (`GT-TMscore`). However, considering that proteins with the same functional descriptions may not necessarily fold into similar structures, we use ProTrek Su et al. (2024b), a tri-modal protein language model, to evaluate textual alignment from both sequence and structural perspectives. Similar

[1]ProtBERT_BFD-512-1e-5-1e-1-text-512-1e-5-1e-1-EBM_NCE-0.1-batch-9-gpu-8-epoch-5

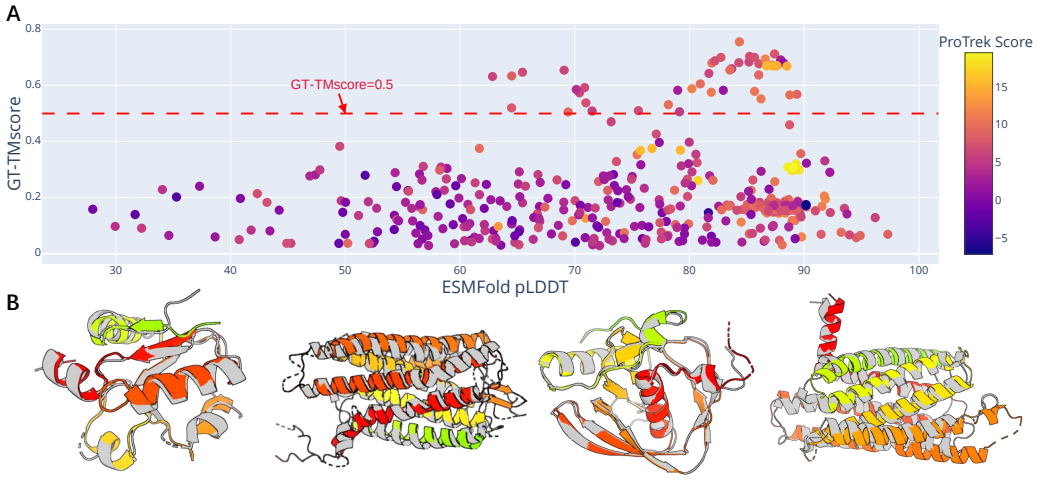

Figure 5: **Generation of out-of-distribution proteins.** **A**: Statistical analysis of proteins designed from unseen compositional descriptions. **B**: Cases where proteins generated from unseen compositional descriptions exhibit strong alignment with ground truth proteins. See Appendix A for a comprehensive list of prompts.

evaluation methods are widely used in the computer vision field (Hessel et al., 2021). Specifically, given a functional descriptions t and a protein p (which can be either protein structure or sequence), we calculate the similarity score, termed as `ProTrek score`, as $s = cos(\tau_t(t), \tau_p(p))$, where $\tau_t(t)$ and $\tau_p(p)$ represent the text and protein encoders, respectively, from ProTrek.

4.2 Generation According to Language Descriptions

To investigate the ability to design *de novo* proteins from long descriptions, the input is constructed as follows: for each protein in the dataset, we gather all of its descriptions, which portray the target protein from various perspectives, and concatenate them into a single sentence. We then randomly select 500 target proteins from the dataset and ask the model to design 5 proteins per sentence, resulting in 500 long descriptions and 2500 designed proteins. To test the robustness of the model to textual input and to facilitate fair comparison with previous methods trained on diverse protein descriptions, the textual inputs are filled into templates different from those used in training or paraphrased by GPT-4.

In Figure 4A, we visualize the assessment of generated proteins from the perspective of textual alignment. The `ProTrek score` is calculated to measure the similarity between the natural language and protein sequence, as well as between natural language and protein structure. All calculated structures are predicted by ESMFold. For each modality (sequence and structure), we report scores in two ways: by taking the maximum and mean scores among the 5 generated proteins from each textual input. Additionally, we calculate the `ProTrek score` and `GT-TMscore` between descriptions and randomly selected proteins from Swiss-Prot for further illustration. We observe that ProteinDT and Chroma can struggle to design proteins that align well with the given text, as their `ProTrek score` and `GT-TMscore` show no significant difference compared to arbitrarily selected proteins from Swiss-Prot. ProGen2-ft, an end-to-end language-to-sequence model, exhibits average performance when generating sequences directly from textual descriptions. This has been explained in Section 2. Pinal outperforms ProGen2-ft, which evidences that the proposed two-stage design approach is more effective than end-to-end training.

We can draw a similar conclusion regarding foldability (Figure 4B). Proteins designed by Pinal exhibit `PAE` and `pLDDT` values comparable to those of proteins from Swiss-Prot, indicating satisfactory structural plausibility. In contrast, sequences designed by ProGen2-ft show varying foldability, suggesting instability and sensitivity to textual input. Lastly, sequences from Chroma and ProteinDT often fail to fold into 3D structures, as indicated by their high `PAE` and low `pLDDT` values.

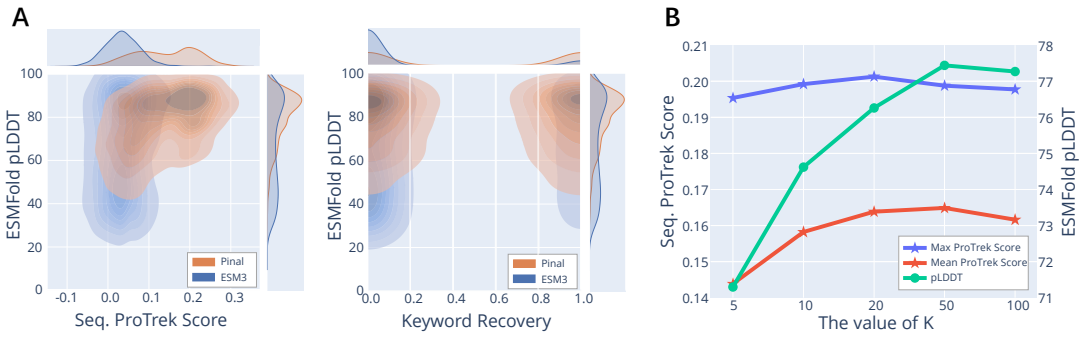

Figure 6: A: Comparing with ESM3. B: Analysis of the optimal value of K.

We further examine the ability to generate protein sequences conditioned on short descriptions, defined as single sentence from individual protein function categories. For each protein function category, we randomly sample 50 sentences from the dataset. Evaluation is conducted by calculating the maximum `ProTrek score` among the 5 sequences and averaging the `pLDDT`. The experimental findings depicted in Figure 4C demonstrate that proteins designed by Pinal exhibit superior alignment with diverse protein descriptions and demonstrate high designability.

To delve deeper into the influence of protein function, we present an evaluation of proteins designed by Pinal across different function categories in Figure 4D. This analysis includes comparison with random `ProTrek Score` and those computed using ground truth proteins for further insight. We observe a positive correlation between foldability and language alignment, where improved language alignment typically corresponds to a greater likelihood of folding into a real protein structure. Pinal excels in designing proteins based on practical descriptions such as protein names and sequence similarity. However, it exhibits limitations when tasked with abstract functional descriptions, such as disruption phenotype and induction. Nevertheless, even with these challenging tasks, Pinal demonstrates the capability to design proteins with high foldability (`pLDDT` > 70).

4.3 Generation Beyond Training Data

Our next focus is on evaluating the generative capacity for proteins that lie beyond the distribution encountered in the training set. Given that proteins in the test set exhibit significant structural differences compared to those in the training set, we will assess this by prompting the model with 100 randomly sampled long descriptions of test proteins. Although our model has seen similar semantics of short sentences during training, it has not been exposed to combinations of these sentences and to proteins similar to those in the test set. We hypothesize that Pinal can design out-of-distribution proteins solely on proper descriptions, which is a significantly more challenging but appealing task.

Although the quality of proteins designed by descriptions from test set falls short compared to those generated by i.i.d. descriptions, the average `pLDDT` remains approximately 70, demonstrating that Pinal can design structurally feasible proteins for novel combinations of functional descriptions (Figure. 5A). Notably, approximately 13% of functional descriptions result in designing proteins with a `GT-TMscore` exceeding 0.5, which indicates the designed proteins and ground truth proteins are closely related or have significant structural homology. Intriguingly, we also find that some of these proteins, despite showing limited similarity to ground truth proteins, exhibit high foldability and textual alignment simultaneously, which may leave for validation through in vitro experiments. In Figure. 5B, we showcase some cases of a designed protein that aligns closely with ground truth.

4.4 Comparative Evaluation with Generative Protein Language Model ESM3

ESM3, as a cutting-edge multimodal protein language model, excels in generating functional proteins from functional keyword. We now evaluate ESM3's ability to design protein sequences solely based on keywords, without partial sequence or structural constraints, and compare its performance with Pinal. Noted that ESM3 recognizes keywords but lacks understanding of natural language. For a fair comparison, we feed Pinal with keywords extracted from InterPro (Paysan-Lafosse et al.,

2023), the dataset ESM3 is trained on. Furthermore, we also report the metric introduced by ESM3, *i.e*. keyword recovery, which is calculated with InterProScan for predicting the consistency of the generated protein with the specified functions.

We notice that Pinal consistently outperformed ESM3 by generating proteins with quality (Figure 6A). The proteins from Pinal exhibit better foldability, as the `pLDDT` of ESM3 designed proteins varied from 20 to 100 evenly. Moreover, Pinal achieves a higher `ProTrek score`, indicating better alignment with desired protein functions. A similar conclusion can be drawn from keyword recovery (see ESM3 Hayes et al. (2024) for details): while half of the proteins from Pinal exhibit predictable functions, only around 10% of the proteins generated by ESM3 do so.

4.5 Design Strategy Ablation

In this section, we investigate the two key design strategy issues that affect the quality of proteins, *i.e*. the importance of designing sequences conditioned on natural language and the optimal value of K, which is introduced in Section 3.1.

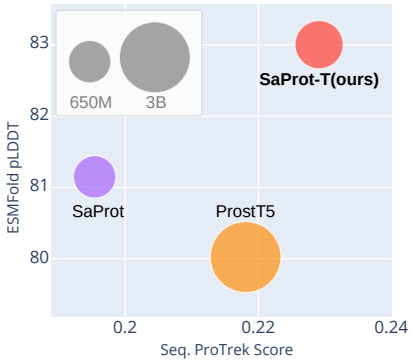

Figure 7: **Sequence Design Comparison.**

The necessity of textual conditioning sequence design. As existing protein language models, *i.e*. SaProt (Su et al.) and ProstT5 (Heinzinger et al., 2023), have shown a strong ability to predict amino acids given foldseek tokens. However, a pertinent question arises: Considering these models' proficiency in decoding structural information, is it essential to incorporate text input conditioning into SaProt's training?

Our answer is yes. Foldseek tokens offer coarse-grained structural cues about desired proteins, allowing for a multitude of possible amino acid sequences. Therefore, accurately predicting amino acids that align well with language descriptions from foldseek tokens is pivotal. To investigate this, we compare the SaProt-T (760M) with vanilla SaProt (650M) and ProstT5 (3B) (Figure 7). We feed them with foldseek tokens derived from natural proteins and evaluate the foldability and textual alignment of the predicted sequences. It is noteworthy that ProstT5, utilizes autoregressive decoding, which is time-consuming. In contrast, SaProt performs inference in a single step to decode all amino acids. Our findings reveal that sequences predicted by SaProt-T, leveraging additional natural language input, outperform in both foldability and textual alignment metrics. This underscores the significance of enhancing SaProt's training with text input conditioning.

The optimal value of K. We next study the number of candidates to explore during the design process. Specifically, we select the top 5 sequences out of K candidates. A larger value of K allows for more extensive exploration but typically increases the time required. Our findings, illustrated in Figure 6B, indicate that compared to the case without our designed optimal sampling strategy (i.e., $K = 5$), there are notable improvements in both the mean `ProTrek score` and `pLDDT` metrics as K increases, up until it reaches 50. Therefore, we selected $K = 50$ as a balanced compromise between exploration efficiency and performance enhancement throughout this paper.

5 Related Work

5.1 *De Novo* Protein Design

De novo protein design refers to the process of creating novel proteins with desired structures and functions from scratch, without relying on existing proteins as templates or starting points in nature (Huang et al., 2016). Current research primarily focuses on two approaches: unconditional design and design conditioned on specific functions, which are roughly categorized into sequence design and structure design.

Most sequence design methods employ the Transformer (Vaswani et al., 2017) architecture to model the design space of proteins. For instance, Progen (Madani et al., 2023; Nijkamp et al., 2023), ProtGPT2 (Ferruz et al., 2022) and RITA (Hesslow et al., 2022) generate amino acid sequence in an autoregressive manner (Radford et al., 2018), while DPLM (Wang et al., 2024c) and EvoDiff (Alamdari et al., 2023) utilize discrete diffusion models (Austin et al., 2021) to generate proteins in a disorderly fashion. These methods are either dedicated to design sequences conditioned on given protein structure (Dauparas et al., 2022; Hsu et al., 2022; Gao et al., 2022), or control tags (Nijkamp et al., 2023; Hayes et al., 2024).

Structure-based design methods often generate novel structures by diffusing in SE(3) space (Yim et al., 2023b). These approaches aim to unconditionally create protein structures suitable for in vitro design (Lin & AlQuraishi, 2023; Lin et al., 2024; Wang et al., 2024b; Yim et al., 2023a; Wu et al., 2024b). Compared to sequence design, direct design in protein structure space offers advantages for tasks like motif scaffolding or binder design that require specific structural features. Therefore, Trippe et al. (2022); Yim et al. (2024); Watson et al. (2023) tailor diffusion models to suit these specific applications. In contrast to relying on inverse folding models to predict protein sequences based on generated structure, (Campbell et al., 2024; Ingraham et al., 2023; Krishna et al., 2024) take a further step by proposing to design structure and sequence simultaneously.

5.2 Alignment Between Language and Protein

To integrate natural language and protein modalities effectively, (Su et al., 2024b; Liu et al., 2023; Xu et al., 2023; Wu et al., 2024a) adopt cross-modal contrastive learning (Radford et al., 2021), enhancing the prediction of protein functions and facilitating bidirectional retrieval between protein and natural language. Simultaneously, the increasing popularity of vision-language models (Liu et al., 2024a) has inspired the training of language models on datasets containing both biological information and natural language. Galactica (Taylor et al., 2023), a pioneering large language model (LLM) in this domain, has been trained on such combined datasets. However, as a general-purpose model, Galactica struggles to provide precise protein descriptions due to limitations in relevant training data. To mitigate this gap, (Zhang et al., 2023; Karim et al., 2022; Pei et al., 2023) focus on integrating meticulously curated biological knowledge into the training of LLMs, with the goal of supporting advancements in biological research. Moreover, (Lv et al., 2024; Abdine et al., 2024; Wang et al., 2024a; Liu et al., 2024b; Guo et al., 2023; Ziegler et al., 2023) specialize in elucidating proteins through tasks such as captioning or answering questions about specific proteins. Conversely, the field of designing proteins from textual descriptions has begun to garner attention among researchers. Recently, two works (Liu et al., 2023; Ingraham et al., 2023) have conducted preliminary explorations with very limited protein description data and a lack of appropriate evaluation metrics, thereby failing to provide comprehensive assessments. In contrast, the proposed Pinal was trained with 30-100 times more text-protein pairs than existing literature, resulting in more aligned protein generation.

6 Discussions

In this paper, we introduce Pinal, a new de novo protein design framework based on natural language. Instead of directly modeling the design space of proteins sequence, we propose a two-stage approach: first, translating protein language descriptions into structural modalities, and then designing protein sequences conditioned on both structure and language. We also developed an elegant optimal sampler to integrate these two stages seamlessly. We conducted a comprehensive dry experiment evaluation, demonstrating that proteins designed by Pinal exhibit high foldability and align well with their natural language prompts, outperforming proteins generated by recent methods. Furthermore, we validate that our method enables generalization beyond the training data, allowing for the design of proteins with unseen functional combinations rather than merely memorizing sequences.

Despite these promising results, several limitations and future research directions warrant exploration:

- While our data provides residue-level protein descriptions (such as mutation effects) (Su et al., 2024b), we primarily use this data to enrich the diversity of functional descriptions

during training rather than for accurate design. This approach is based on two key considerations. Firstly, residue-level information varies widely across different proteins, posing challenges in developing a universal model for understanding the function of each amino acid and subsequently generating new proteins. Relying solely on the current data is insufficient for establishing a robust model. Secondly, the absence of reliable methods to predict amino acid functions from any protein sequence limits our ability to validate the efficacy of designed proteins.

- Although the datasets we utilized are considered superior within the biological community, they are still incomparable, both in terms of quality and quantity, to datasets in other fields such as natural language processing or visual generation. A dilemma arises because proteins that biologists are interested in often involve complex reaction mechanisms but lack related descriptions, making them challenging for models to design. Conversely, proteins with ample descriptions are easier for models to handle but have already been extensively studied and may not be as critical. This necessitates models to possess strong generalization abilities, underscoring the importance of accessing higher-quality datasets.
- In this paper, we limit our evaluation to dry experiments. In the future, we plan to validate the designed proteins through wet lab experiments.

REFERENCES

Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: Multimodal protein's function generation with gnns and transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Atomic-Level Accuracy. Design of a novel globular protein fold with. *science*, 1089427(1364):302, 2003.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, pp. 2023–09, 2023.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31 (1):365–370, 2003.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338 (6110):1042–1046, 2012. doi: 10.1126/science.1219021. URL https://www.science.org/doi/abs/10.1126/science.1219021.

Michael J Dougherty and Frances H Arnold. Directed evolution: new parts and optimized function. *Current opinion in biotechnology*, 20(4):486–491, 2009.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Zhangyang Gao, Chen Tan, and Stan Z Li. Foldtoken3: Fold structures worth 256 words or less. *bioRxiv*, pp. 2024–07, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Han Guo, Mingjia Huo, Ruiyi Zhang, and Pengtao Xie. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. *Authorea Preprints*, 2023.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Martin Steinegger, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. biorxiv. *bioRxiv*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Daniel Hesslow, Niccoló Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Md Rezaul Karim, Hussain Ali, Prinon Das, Mohamed Abdelwaheb, and Stefan Decker. Question answering over biological knowledge graph via amazon alexa. *arXiv preprint arXiv:2210.06040*, 2022.

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Tim Kucera, Carlos Oliver, Dexiong Chen, and Karsten Borgwardt. Proteinshake: building datasets and benchmarks for deep learning on protein structures. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiaohan Lin, Zhenyu Chen, Yanheng Li, Xingyu Lu, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao, and Jun Zhang. Protokens: A machine-learned language for compact and informative encoding of protein 3d structures. *bioRxiv*, pp. 2023–11, 2023.

Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. *arXiv preprint arXiv:2301.12485*, 2023.

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv preprint arXiv:2405.15489*, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, et al. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. Prott3: Protein-to-text generation for text-based protein understanding. *arXiv preprint arXiv:2405.12564*, 2024b.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*, 2024.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2310.07276*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*.

Jin Su, Zhikai Li, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, Dacheng Ma, The OPMC, Sergey Ovchinnikov, and Fajie Yuan. Saprothub: Making protein modeling accessible to all biologists. *bioRxiv*, pp. 2024–05, 2024a.

Jin Su, Xibin Zhou, Xuting Zhang, and Fajie Yuan. Protrek: Navigating the protein universe through tri-modal contrastive learning. *bioRxiv*, pp. 2024–05, 2024b.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. arxiv 2022. *arXiv preprint arXiv:2211.09085*, 10, 2023.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pp. 2022–02, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chao Wang, Hehe Fan, Ruijie Quan, and Yi Yang. Protchatgpt: Towards understanding proteins with large language models. *arXiv preprint arXiv:2402.09649*, 2024a.

Chentong Wang, Yannan Qu, Zhangzhi Peng, Yukai Wang, Hongli Zhu, Dachuan Chen, and Longxing Cao. Proteus: exploring protein structure generation for enhanced designability and efficiency. *bioRxiv*, pp. 2024–02, 2024b.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024c.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Kevin E Wu, Howard Chang, and James Zou. Proteinclip: enhancing protein language models with natural language. *bioRxiv*, pp. 2024–05, 2024a.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024b.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. Protst: Multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning*, pp. 38749–38767. PMLR, 2023.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023a.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023b.

Jason Yim, Andrew Campbell, Emile Mathieu, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Frank Noé, et al. Improved motif-scaffolding with se (3) flow matching. *ArXiv*, 2024.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Cheyenne Ziegler, Jonathan Martin, Claude Sinner, and Faruck Morcos. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14(1): 2222, 2023.

A Showcase Prompts

Prompts for Fig. 1:

1. The primary role of this protein is to facilitate the reaction: beta-D-fructose 1,6-bisphosphate + H2O = beta-D-fructose 6-phosphate + phosphate through its enzymatic function.
2. Catalyzes the formation of 5-methyl-uridine at position 1939 (m5U1939) in 23S rRNA.
3. The protein can be found in Secreted. Compound that limits ion channel opening The protein sequence derives from the organism named Brush-footed trapdoor spider. The protein sequence derives from the organism named Trittame loki. The GO term encompassing toxin activity involves molecular function for this protein. Is included in the neurotoxin 14 (magi-1) family The subfamily labeled as 03 (ICK-30-40) sodium channel inhibitor activity falls under the GO term associated with this protein in relation to molecular function. U17-BATX-Tl1a is the official name of this protein. The GO term encompassing extracellular region involves cellular component for this protein. The assigned designation for this protein is U17-barytoxin-Tl1a.
4. The GO term for this protein constitutes membrane when considering cellular component. Is a member of the syntaxin protein family cellular component is involved in the GO term associated with this protein, encompassing presynapse. This particular protein can be found at Membrane. Golgi trans cisterna is encompassed by the GO term associated with this protein regarding cellular component. The GO term for this protein includes synaptic vesicle docking concerning biological process. The name Caenorhabditis elegans corresponds to the organism from which the protein sequence originates. In regards to molecular function, the GO term associated with this protein integrates SNAP receptor activity. The promotion of transport vesicle movement to target membranes is aided by SNARE (By similarity) Possibly operates in retrograde trafficking and endocytic recycling pathway (By similarity) This is the protein's designated term, Putative syntaxin 6.
5. The metabolic pathway associated with this protein encompasses Organic acid metabolism; propanoate degradation as well.
6. The GO term for this protein includes regulation of DNA repair concerning biological process.
7. Biological process is involved in the GO term associated with this protein, encompassing phagosome-lysosome fusion.
8. Carbohydrate biosynthesis; gluconeogenesis is an integral part of the metabolic pathway associated with this protein.
9. The presence of this protein enables the reaction: ATP + L-threonyl-[protein] = ADP + H(+) + O-phospho-L-threonyl-[protein] to be catalyzed through its enzymatic activity.
10. The catalytic activity of this protein allows for the reaction: guanosine(46) in tRNA + S-adenosyl-L-methionine = N(7)-methylguanosine(46) in tRNA + S-adenosyl-L-homocysteine to proceed.
11. Catalyzes the conversion of 3-phosphate to a 2',3'-cyclic phosphodiester at the end of RNA The enzyme's mechanism involves a three-step process: (A) adenylation by ATP, (B) transfer of adenylate to an RNA-N3'P to yield RNA-N3'PP5'A, and (C) initiating a reaction with the adjacent 2'-hydroxyl on the 3'-phosphorus in the diester linkage to generate the cyclic end product. The biological role of this enzyme is unknown but it is likely to function in some aspects of cellular RNA processing.
12. The organism associated with the protein sequence is referred to as Streptomyces cinnamonensis. The GO term for this protein constitutes antibiotic biosynthetic process when considering biological process. The metabolic pathway associated with this protein incorporates Antifungal biosynthesis; monensin biosynthesis. The protein sequence is attributed to the organism Streptomyces virginiae. In the taxonomic hierarchy, the source organism of this protein falls into the category Streptomyces. The designated name for this protein is ORF4. Is needed for correct cyclization of the oligoketide leading to isochromanequinone formation This protein is designated as Granaticin polyketide synthase bifunctional cyclase/dehydratase.

Prompts for Fig. 5:

1. Tstd3 is the official gene name for this protein. The proteome, comprising Chromosome 4, contains this protein. This protein is designated as Thiosulfate sulfurtransferase/rhodanese-like domain-containing protein 3. This protein's source organism belongs to the category Euteleostomi in the taxonomic hierarchy. Mouse denotes the organism that yields the protein sequence. The name Mus musculus identifies the organism that provides the protein sequence. The specific term assigned to this protein is Rhodanese domain-containing protein 3. thiosulfate sulfurtransferase activity is encompassed by the GO term associated with this protein regarding molecular function.
2. The specific gene marker srd-48 is indicative of this protein's presence. This protein's source organism belongs to the category Rhabditidae in the taxonomic hierarchy. Caenorhabditis elegans is the name given to the organism that produces the protein sequence. This is the protein's designated term, Protein srd-48. This protein can be situated at Membrane. Belongs to the nematode receptor-like protein srd family The proteome, comprising Chromosome X, contains this protein. The GO term of this protein covers membrane with respect to cellular component. Serpentine receptor class delta-48 is the official designation for this protein. The ORF Name F17A2.9 denotes the gene associated with this protein.
3. This protein's cofactor activity is enabled by Fe(2+). Belongs to the polypeptide deformylase family [Detaches the N-terminal methionine of newly synthesized proteins from its formyl group] Efficient rates of reaction hinge on the presence of at least a dipeptide. Although N-terminal L-methionine is essential for the enzyme's activity, it exhibits broad specificity at other positions. This is the protein's designated term, Polypeptide deformylase. The protein is recognized by the term PDF. Affinity for a Fe(2+) ion. In regards to biological process, the GO term associated with this protein integrates translation. The proteome, comprising Chromosome, contains this protein. metal ion binding is encompassed by the GO term associated with this protein regarding molecular function. The organism responsible for the protein sequence is denoted as Symbiobacterium thermophilum (strain T / IAM 14863).
3. Membrane is the position of this specific protein. This protein is designated as Protein srd-2. This protein is identified by the gene signature srd-2. Belongs to the nematode receptor-like protein srd family The gene responsible for this protein is specified by the ORF Name R05H5.1. This protein's source organism is classified under Rhabditomorpha in the taxonomic hierarchy. The assigned designation for this protein is Serpentine receptor class delta-2. The GO term of this protein covers membrane with respect to cellular component. The consensus is that it acts as a chemosensory receptor The organism associated with the protein sequence is referred to as Caenorhabditis elegans.